WindWard environmental LLC

200 West Mercer St. ♦ Suite 401 ♦ Seattle, WA 98119
Phone: 206.378.1364 ♦ Fax: 206.217.0089 ♦ www.windwardenv.com

MEMORANDUM

To:	EPA/LWG Ecorisk Team
From:	Lisa Saban, Helle Andersen, Mike Johns, Lorraine Read, Teresa Michelsen
Subject:	Summary of November 21, 2005 Benthic Meeting
Date:	December 6, 2005

Attendants in person or on the phone: Jay Field, Teresa Michelsen, Lorraine Read, Mike Johns, Lisa Saban, Rick Applegate, Rob Pastorok, Nancy Musgrove, Joe Goulet, Jennifer Peterson, Mike Anderson, Ben Shore, Taku Fuji, Chris Thomson, Helle Andersen.

Hand-outs: Interpretive maps of chemistry and bioassay data based on no-hit designations and the Floating Percentile Model. Outline of the Benthic Interpretive Report. *I didn't receive an outline of the benthic interpretive approach. If this was the presentation, it went too fast to comment on, and I did not get an electronic copy after the fact. They should submit this, though.*

Short summary of meeting agenda:

- The meeting started out with presentations of the Logistic Regression Model by Lorraine Read and the Floating Percentile Model by Teresa Michelsen. As an introduction to the two models Helle Andersen gave a short presentation of manipulation and reduction of the chemical data which was done before the work with the models were initiated. *It would be good to have this summary written down somewhere because am still not clear on all they did to manipulate and reduce the chemical data, and if it was appropriate.*
- After the presentations the letter from EPA regarding the benthic interpretive approach (October 26, 2005) was discussed by going over each of the main issues listed in the letter.
- Lorraine Read presented her findings with the Logistic Regression Model and other analyses (scatter plots and more) she had performed on the data.

- Jay Field gave a presentation on his findings with the Logistic Regression Model and pointed out issues including the good correlation between % fines and toxicity that need further discussion and evaluation. The interpretation of this finding from a chemistry benthic risk will need to be discussed within the risk assessment/risk management framework.
- The interpretive maps of chemistry and bioassay data based on no-hit designations and the Floating Percentile Model were discussed including an approach that used the data to screen out the areas showing no toxicity.
- At the end of the day Mike Anderson gave a summary of his findings while working with the Floating Percentile Model. *Mike discussed some of the problems with the model, some of which still need to be resolved. Teresa's responses to his detailed memo were not received prior to the meeting, but were handed to him in hard copy at the meeting. He is going over this info and will respond.*
- The team is finding a link between conventional parameters and toxicity response-both teams will investigation this further.

Agreements reached concerning issues raised in the EPA letter regarding the interpretive report:

- The Logistic Regression Model will use pooled endpoints both by species (*Hyalella* growth and mortality, and *Chironomus* growth and mortality) and an overall pooled endpoint (*Hyalella* and *Chironomus* growth and mortality). Questions were raised regarding the appropriateness of pooling when one or more individual endpoints did not have a good model fit. The benthic report will explore this evaluation in more detail (pooling vs non-pooling).
- The Floating Percentile Model will use three individual endpoints. One of the four endpoints (*Hyalella* growth) will most likely not be included in the model due to poor performance. The poor performance will be discussed in the benthic interpretive report (see action items). A section will be included in the benthic interpretive report discussing the effect of using pooled endpoints similar to the Logistic Regression Model and provide examples of the results. However, because the model gives better results using each endpoint individually, the final model run will be completed using the three individual endpoints. To address the effects of pooling the endpoints, maps will be provided showing the outcome of the three separate endpoints in a pooled format (probably in a "pie chart per station" format). *They presented this as a proposal at the meeting, but I wasn't aware we agreed to it. I would recommend including Hyalella growth as an endpoint in the report (pooled endpoints both by species and overall). We can then make a decision on how to interpret the results. They seem to want to evaluate each endpoint separately, which we have some problems with (e.g. between there may be some confounding effects*

between in growth where there is high mortality). I guess this is fine as long as they also evaluate pooled endpoints between species and overall pooled (Hyalella and Chironomus growth and mortality).

- Control-normalization was found to be a minor issue. The growth endpoints were already normalized as outlined in the Benthic Approach memo. The control performance for the mortality endpoint was very good in all the bioassays and the normalization process would therefore not make much difference. Because the issue was raised by EPA late in the process, it was agreed that in the effort of not loosing time and effort already spent, the control-normalization would not be done for the mortality endpoint. *We asked for control normalization in our memo, and it would be better to do it this way. They resisted, and we approved? Anyway, it may make a difference for those samples at the low tox end (maybe those that were not stat. distinguishable). Therefore, if they don't want to do it they should show that it does, in fact, not make a difference by presenting it both ways. That way we can agree with their analysis.*

- The interpretive report will include three hit/no-hit toxicity thresholds (level 1, 2, and 3).

  For the Logistic Regression Model these will be:

  1) <90% control-normalized growth and survival
  2) <80% control-normalized growth and survival
  3) <70% control-normalized growth and survival

  For the Floating Percentile Model these will be:

  1) <90% survival or <90% control-normalized growth *(if they can go back and add this one in, then why can't they do the 70 and the 80 as well?) I am not sure with the FPM that adding an additional level at the lowest end is more beneficial than looking at slicing the magnitude of toxicity up at the 20 and 30 area where you would see more effects (esp. those moderate in nature).*
  2) SQS definitions (see benthic approach memo)
  3) CSL definitions (see benthic approach memo)

  The thresholds for the Logistic Regression Model were selected to parallel the work Jay Field is doing with the data, whereas the thresholds for the Floating Percentile Model were selected to follow regional guidelines. *I wasn't under the impression we agreed to do this for sure, we agreed Mike Anderson would do the analysis. As for what threshold level we would select has yet to be determined.* The work that has been done to date with the two sets of thresholds shows that there is not much difference in the outcome of the two models. A section in the interpretive report will summarize the differences.

- A list of chemicals detected in sediment but not included in the models will be provided in the interpretive report. The chemicals excluded from the models were based on less than 30 detected concentrations. In the discussion Jay Field pointed out that he uses a cut-off value of less than 100 detected concentrations.
- The interpretive report will include a definition of the "N" qualifier and a summary statement regarding how many data points were excluded because of this qualifier. *We should internally have a position on if it is appropriate to exclude "N" qualified data. This is for contaminants where there was presumptive evidence for an analyte, and for metals the analysis was not in control limits. For organics this was where the analyte exhibited low spectral parameters. What does this mean? Does anyone understand this qualifier.*

Action items:

- Both teams will plot the stations where TPH was analyzed to verify that the stations were selected based on sources.
- More work will be done to evaluate the effects of grain size. Teresa Michelsen will combine the *Hyalella* growth endpoint with percent fines and see if this improves the performance of this endpoint. If the endpoint is improved, it will be included in the model and the benthic interpretive report. *It should be included regardless.*
- Porewater ammonia data from the bioassay should be compared with the ammonia concentrations measured in the bulk sediment to evaluate if there is any correlation.
- Teresa Michelsen and Mike Anderson will continue working together on issues related to the Floating Percentile Model.
- The benthic interpretive report will be submitted to EPA in early February, 2006.
- A phone conference meeting may be held prior to submittal of the benthic interpretive report to discuss any issues that may come up during the final runs of the two models.

Jay’s comments:

1) FPModel endpoints: I do not think we agreed that the FPM should use only the individual endpoints. We certainly did not agree that the FPM should ignore the Hyalella growth endpoint. Poor model performance is not a good reason to ignore the most sensitive endpoint. As we pointed out in the July meeting, in the memo to LDW, and again in the 11/21 meeting, the pooled growth/survival results for the each species is a better way to look at the growth results because growth is not independent of survival. If they insist on using the 4 individual endpoints, I think it is reasonable to ask them to apply the FPM to the pooled results for each species as well.

2) control-normalized results: I still do not agree with Teresa's approach to control-normalization (subtracting the control result from the test result rather than dividing test by control as is commonly done). She also appears to be using this approach for growth (which I think is different than in their original benthic methods memo). Teresa is following (creating) her own precedent. However, I do agree that the results for this data set are likely to be minimal.

3) N-qualified data: attached is a spreadsheet summarizing (min, max, and number of samples) the N-qualified results for each chemical. Note that there are some high concentrations that would be excluded for a variety of chemicals. I have not been excluding them in my analyses (I apparently missed the discussion of this qualifier).

4) summing chemicals: I am using total PCBs. Because of the strong correlation between individual PAHs and LPAH and HPAH, I do not think it's a significant problem to use LPAH & HPAH instead of the individual PAHs. I have not looked at the relationship between total DDTs and the six isomers. If they use the summed values for PAHs and DDTs, they should show the basis in the report (consistent composition throughout the study area).

5) in my view, the most critical issue at this time is to resolve Mike A's issues on the FPM, to make sure that the final model results are reproducible.

Jenn’s comments:

Technical Issues:

1. Summing contaminant classes versus using individual contaminants in the model: The proposal from the LWG was to sum DDTs, PAHs and PCBs (I think that was all). However, Mike and Jay are currently not summing. It seems better o sum at a later stage. For example, with the FPM Mike

is running, you can see if contaminants are correlated and maybe should be summed from the results of the analysis. If contaminants are co-varying they will show this by where they "float" in the analysis". If this is shown, summing at that point makes sense, but maybe not before. It would be better to present unsummed analysis and summed - that way we have the information we need to make a decision on what numbers are more appropriate.

2. Alpha levels: In the meeting they mentioned they were running the analysis using an alpha level of 0.05. The alpha levels represent the probability of making incorrect conclusions that the treated sample is toxic or contains chemical residues not found in the control or reference sample (Type 1 error). By setting this probability low (0.05), the likelihood that one erroneously concludes there are no differences among the mean responses in the treatment, control or reference samples (Type 2 error) increases (low power). Type 2 errors would lead to conclusions that the sample is not toxic (or different from control or reference), when in fact there is a difference. Type 2 errors are important to minimize in environmental investigations, since, if left undetected, these errors can lead to continued short- and long-term effects (ASTM 2003; EPA 2000a). In order to avoid this, an alpha of 0.1 can be used (and is in the work plan), which would increase the power of the test and the probability of detecting a reduction relative to the control mean. They are currently eliminating some samples on the basis that they are indeterminate in difference from the control at an alpha of 0.05 (I think from the meeting there were about 11 eliminated). However, they may be determinate at an alpha of 0.1. These low responses may be important in the model - especially the FPM. In the work plan they state "if the analysis of the toxicity test data finds that the power for the data set is low, the alpha level may be raised to 0.1 as suggested in ASTM guidelines (2003)." From the meeting there was not mention they were moving forward with that analysis, however, I would recommend the report should include the analysis at an alpha of 0.1 and indicates how this changes the conclusions.

3. What contaminants should be eliminated from the model: This relates to removing contaminants on the basis that they are not drivers of toxicity (e.g. aluminum). However, Mike A's analysis showed that some were slight predictors of toxicity. It may still be removed later on the basis that it is not a toxicity driver, but the report, (and their analysis) should include these contaminants (see "3" below). The analysis (at least for the FPM) will clearly show contaminants that aren't driving toxicity, and this will provide justification for dropping contaminants.

4. The results of the bioassay tests and modeling effort may show that additional lines of evidence may be important in interpreting the bioassay results (e.g. EqP or pore water testing).

Larger Issues Include (may need more manager input):

1. Running the FPM - there are still discrepancies between Teresa and Mike's models that must be resolved at a fundamental level. We don't want to be dealing with problems in replicating the FPM further down the line when we are also having to analyze results. I would recommend that these issues be worked out prior to submittal of the report, but more

importantly that ALL steps she takes to get the FPM values be explicitly written out for each chemical / decision made. This should be at the detail that someone reading the report can replicate what was done.

2. Discrepancies between the FPM and the logistic regression results: PAHs are a good example of this. The FPM method is calculating very high dry weight concentrations of PAH threshold numbers using this method that the government team does not agree with (and Jay has said is a non-starter).

2. What endpoints should we be considering? The Hyalella growth endpoint appears to be producing different results than the other test endpoints. Teresa wants to remove this from her analysis because it is not producing reliable results, even though it is being used as a part of the logistic regression modeling. I don't think the team members agree with this assessment. I would recommend model runs for this endpoint should be included in the report, along with pooled endpoint runs that include this endpoint. We can then assess what it means after we see the data.

3. What do we want the models to do? Loraine brought up this point and it is a very good one. Do we want the model to provide information on the chemicals detected in Portland Harbor or find the most predictive component that is predictive of toxicity (e.g. even if it is a conventional parameter)? You can run the models and get numbers for each chemical - if it is not contributing to toxicity this number will most likely be the AET from the dataset. However, I think this is useful information to anyone reviewing this report. I would recommend that most chemicals be run in order to justify their removal (which is easy running the FPM, but maybe not the logistic regression). Mike Anderson did this very quickly, and showed that some chemicals were not contributing to toxicity on the basis of the analysis. Numbers behaving in this manner were flagged with an AET value. By doing this it is easy to see that contaminant X wasn't a driver for toxicity at the highest detected concentration of X. This information is useful. The alternative is to find the most predictive indicator of toxicity, which may be a conventional parameter such as bulk ammonia, bulk sulfide or percent fines, or it may include a very limited list of contaminants. The downside here is that this approach may provide limited data on a wider list COPCs. If we go this route, bioassays to validate the model should definitely be done, and realize that it will not translate easily into cleanup numbers.

4. What hit/no hit thresholds should we be considering? We gave some direction in our memo to them. However, they resisted going to the same thresholds between methods (for the FPM) in order to comply with consistency with other programs (which is odd because the "other programs" are still Teresa's work, but for Washington State). We had originally proposed using 10, 20 and 30 (or 90, 80 and 70) to correspond with NOAA's levels. Teresa did stat only, 10 and 25 for Washington State. Therefore, we got pushback on using the NOAA thresholds for Teresa's FPM analysis. Jay seems to think this is o.k. because the threshold levels don't matter too much as long as you get information at several levels for the model. I agree with him for the logistic regression model (because eventually you are developing a continuous model for which you can pick anywhere on the curve to correspond with magnitude of toxicity and prob of toxicity [jay correct me if I am wrong] for use in management objectives), but this is not the case for the FPM. Magnitude of toxicity (hit/no hit) levels need to be selected

before hand and that is all the data you will have to make decisions. You can't for example select another threshold (e.g. something between the 10 and the 25) without re-running the analysis because you do not have a continuous distribution like the logistic regression model. We concluded that because of the resistance and since Mike had the data he could run the 10, 20, and 30 for the government team and we could analyze any differences between the different levels of magnitude of toxicity. However, it would have been better to stay consistent, and I think the three levels indicating magnitude of toxicity would have been helpful in interpreting the data for the FPM.